Meaningful Data Compression and Reduction of High-Throughput Sequencing Data

RUTGERS, THE STATE UNIVERSITY OF NEW JERSEY

PI: SCHLIEP, ALEXANDER

DESCRIPTION (provided by applicant): High-throughput sequencing (HTS), a technology to unravel DNA sequences on a large scale, is pervasive in clinical and biological applications such as studying the spectrum of genetic variations and their relation to disease. Due to further reductions in cost, sequencing is expected to gain significant momentum, since it will replace commonly used genetic tests in clinical care for life-threatening diseases such as cancer, and consequently produce enormous amounts of data. The rise of personalized medicine will eventually lead to the point where every individual can be routinely screened for genetic risk factors using HTS. The goal of the proposed research is to boost the analysis of HTS data with a compressive genomics middle-ware that provides compressed reduced representations of HTS data. The representations are meaningful in that sequence information which is likely to cover the same genomic location in the sequenced genome will be brought together. As existing and future methods and algorithms can operate directly on this representation, the proposal not only realizes a saving in space and transmission times, but also in CPU time needed for analysis. The project has three aims: 1) Develop a clustering algorithm for single and paired HTS read libraries that rapidly recognized overlapping. Establish a lossless compression scheme based on clusters, which facilitates downstream computations directly on the compressed data without decompression. Extend the approach to joint compression of multiple HTS libraries. 2) Introduce meaningful reduced representations which further decrease memory demands by prioritizing sequence information likely to be correct and discarding information likely to be erroneous. 3) Adapt important HTS analysis tools to our compressive genomics approach, in particular read mapping, de novo genome assembly by using cluster consensus sequences as virtual, elongated reads for a hybrid assembly scheme, and discovery of structural variants based on cluster mapping positions and ambiguities in assignment of sequences to clusters. Our results will aid in improving health care outcomes by increasing analysis quality, lowering costs and making the analysis of HTS data more widely accessible. This will impact areas of scientific inquiry from understanding genetic variations underlying disease to personal genomics. PUBLIC HEALTH RELEVANCE: High-throughput sequencing (HTS), a technology to unravel gene sequences on a large scale, is pervasive in clinical and biological applications such as cancer research, and is expected to gain enormous momentum in future personalized medicine applications. Storage, transfer and analysis of the resulting, very large data sets poses substantial problems for research centers and individual researchers. We will develop meaningful compressed and reduced representations, which not only reduce the size of the data, but greatly accelerate and improve analysis through "compressive genomics", thus building a path into the era of personal genomics.